Thompson *et al.*
BMC Medical Research Methodology (2022) 22:222
https://doi.org/10.1186/s12874-022-01699-2

BMC Medical Research Methodology

RESEARCH

Open Access

Cluster randomised trials with a binary outcome and a small number of clusters: comparison of individual and cluster level analysis method

Jennifer A. Thompson[1*], Clemence Leyrat[2], Katherine L. Fielding[1] and Richard J. Hayes[1]

Abstract

Background: Cluster randomised trials (CRTs) are often designed with a small number of clusters, but it is not clear which analysis methods are optimal when the outcome is binary. This simulation study aimed to determine (i) whether cluster-level analysis (CL), generalised linear mixed models (GLMM), and generalised estimating equations with sandwich variance (GEE) approaches maintain acceptable type-one error including the impact of non-normality of cluster effects and low prevalence, and if so (ii) which methods have the greatest power. We simulated CRTs with 8–30 clusters, altering the cluster-size, outcome prevalence, intracluster correlation coefficient, and cluster effect distribution. We analysed each dataset with weighted and unweighted CL; GLMM with adaptive quadrature and restricted pseudolikelihood; GEE with Kauermann-and-Carroll and Fay-and-Graubard sandwich variance using independent and exchangeable working correlation matrices. P-values were from a t-distribution with degrees of freedom (DoF) as clusters minus cluster-level parameters; GLMM pseudolikelihood also used Satterthwaite and Kenward-Roger DoF.

Results: Unweighted CL, GLMM pseudolikelihood, and Fay-and-Graubard GEE with independent or exchangeable working correlation matrix controlled type-one error in > 97% scenarios with clusters minus parameters DoF. Cluster-effect distribution and prevalence of outcome did not usually affect analysis method performance. GEE had the least power. With 20–30 clusters, GLMM had greater power than CL with varying cluster-size but similar power otherwise; with fewer clusters, GLMM had lower power with common cluster-size, similar power with medium variation, and greater power with large variation in cluster-size.

Conclusion: We recommend that CRTs with ≤ 30 clusters and a binary outcome use an unweighted CL or restricted pseudolikelihood GLMM both with DoF clusters minus cluster-level parameters.

Keywords: Cluster-level analysis, Cluster level analysis, Generalised linear mixed model, Generalised estimating equations, Comparison of methods, Cluster randomised trial, Small number of clusters

Background

Cluster randomised trials (CRTs) are often designed with a small number of clusters [1], but it is not clear which analysis methods are optimal when the outcome is binary.

In a CRT, groups of individuals known as clusters, such as health clinics or villages, are randomly assigned

*Correspondence: Jennifer.Thompson@lshtm.ac.uk

[1] Department of Infectious Disease, London School of Hygiene & Tropical Medicine, London, UK
Full list of author information is available at the end of the article

to receive either a control or intervention condition. Observations from the same clusters are likely to be more similar to one another than observations from different clusters, and it is well known that this correlation needs to be taken into account to prevent confidence intervals that are too narrow and p-values that are too small [2].

There are three broad types of analysis that can be used for CRTs: cluster-level analyses, generalised linear mixed effect models (GLMM), and generalised estimating equations with sandwich standard errors (GEE). In a cluster-level analysis, the observations from each cluster are summarised, and these cluster-level summaries are analysed using simple methods for independent data, most commonly with a weighted or unweighted t-test. In a GLMM, the correlation of observations in the same cluster is directly modelled by including a random effect for cluster. GEE assume a working correlation structure for observations in the same cluster, and standard errors are calculated allowing for the observed correlations in the data.

There is a wealth of literature on each type of analysis method with a small number of clusters. Cluster-level analysis is known to maintain control of type-one error with a small number of clusters and non-normally distributed outcomes [3, 4]. GLMM or GEE require small sample corrections with a small number of clusters to maintain an acceptable type-one error rate [5–8]. GLMM requires use of restricted maximum likelihood and comparison of test statistics to a t-distribution rather than the normal distribution [5]. GEE require use of a bias corrected standard error as well as use of a t-distribution. There have been many bias corrected standard errors developed and each performs well in different scenarios. Some are known to be conservative in scenarios common to CRTs [9], while others have closer to nominal type one error [7, 8, 10]. Continuous outcomes and binary outcomes with a high prevalence are well studied [6, 10, 11], and to date, all assessments have assumed model assumptions are met. Cluster-level analysis is known to maintain control of type-one error with non-normally distributed outcomes [3, 4]. GLMM are robust to some degree of non-normality [12], but this has not been explored for a small number of clusters or for GEE. With a large number of clusters, individual-level analysis with GLMM or GEE has greater power than a cluster-level analysis with varying cluster size [2]. Power is known to be reduced by use of the GEE small sample corrections with continuous outcomes [11], but this has not been studied for a binary outcome. We provide more detail of this previous literature for each type of analysis in the "Background to methods" section below.

Binary outcomes are the most common type of outcome for CRTs [1], but raise problems for the CRT analysis methods. Cluster-level methods become more challenging when some clusters have no events of interest, and GLMM require numerical methods of integration of the random effects [2]. Commonly used effect measures such as the odds ratio are also non-collapsible, so that GEE estimates have a different interpretation to cluster-level analysis and GLMM.

In this paper, we address some remaining gaps in the literature for binary outcomes: Is it possible to control type-one error for each method with a low prevalence outcome? If type-one error can be controlled, which type of method has greatest power? How sensitive is each type of analysis to non-normality of cluster effects? We begin by describing the analysis methods in more detail and reviewing previous literature on use of these methods with a small number of clusters for the estimation of an odds ratio. We then report an extensive simulation study that addresses our three research questions. We demonstrate the impact of analysis choice on an illustrative CRT and provide recommendations to trialists.

Methods

Background to analysis methods

In this section, we will review the analysis methods that have been shown to maintain nominal type-one error for CRTs with a binary outcome and a small number of clusters. We only consider analyses that do no adjust for covariates.

The estimates from these methods have different interpretations. The cluster-level analysis and GLMM provide cluster-specific intervention effect estimates; these are estimates of the average effect comparing one participant given the intervention to one participant given the control drawn from the same cluster. GEE provide population-average (also known as marginal) intervention effect estimates; these are estimates of the average effect comparing one participant given the intervention to one participant given the control drawn from the population.

We consider a CRT with n clusters randomised to either an intervention or control condition with outcomes $y_{ijk} = 0, 1$ in arm i in cluster j for individual $k = 1, \ldots, m_{ij}$ where m_{ij} is the number of observations in arm i in cluster j.

Cluster-level analysis

In a cluster-level analysis, individual observations are summarised for each cluster. With a binary outcome, a proportion is commonly used, but for comparability with GLMM and GEE, here we will consider the log odds of the outcome in each cluster, so that we estimate an odds ratio intervention effect. In clusters with no events of interest, the log odds are not defined. To avoid this, we added 0.5 events and 0.5 non-events to each cluster [2], so that the cluster log odds are defined as:

$$s_{ij} = \frac{\sum_{k=1}^{m_{ij}} y_{ijk} + 0.5}{m_{ij} - \sum_{k=1}^{m_{ij}} y_{ijk} + 0.5}$$

These cluster log-odds are independent of one another, so simple analysis procedures can then be used to derive a confidence interval and p-value. A t-test with degrees of freedom as n – 2 has been shown to maintain type-one error with as few as 6 clusters [13].

The t-test is known to be robust to relatively large deviations from the assumption of normality of cluster summaries [14]; supporting Fig. 1A and C show examples of non-normally distributed outcomes that do not affect t-test performance. However, the method becomes inefficient if clusters vary in size [15]. To improve the efficiency, several weights for the clusters and use of a weighted t-test have been suggested, but the performance of these weights remains unclear. While weighting clusters could maximise the use of information, this has to be considered against the uncertainty in the estimation of the weights themselves [2].

Here, we compare the performance of the following cluster weights:

Unweighted analysis All clusters are given the same weight.

Inverse-variance weights Inverse variance weights account for the information provided by each cluster by weighting each cluster inversely to the variance of the cluster summary. From a Taylor series expansion, the variance of the cluster log-odds is approximated by

$$Var\left(\log\left[\frac{\widehat{p}_{ij}}{1-\widehat{p}_{ij}}\right]\right) \approx \sigma_b^2 + E\left[\frac{1}{p_{ij}\left(1-p_{ij}\right)m_{ij}}\right]$$

where $\widehat{p}_{ij}$ is the observed proportion of observations with the outcome in arm i in cluster j, σ_b^2 is the variance of the true cluster log-odds, p_{ij} is the true proportion of observations with the outcome in arm i in cluster j.

We can substitute the intracluster correlation coefficient (ICC) into this formula. The ICC is the between cluster variability divided by the total variability. We use the definition for the ICC on the log odds scale of $\rho_i = \sigma_b^2/\left(\sigma_b^2 + 1/p_i[1-p_i]\right)$ [16] where p_i is the prevalence of the outcome in each arm.The variance of the cluster log-odds leads to different inverse-variance weights depending on assumptions about the prevalence of the outcome within clusters. Under a null hypothesis of the same mean prevalence in both arms, then the ICC is the same for both arms so that $\rho_0 = \rho_1 = \rho$, and we get inverse-variance weights suggested by Kerry and Bland [15] (see supplementary text for derivation)

$$w_{ij} = \frac{m_{ij}}{1 + \left(m_{ij} - 1\right)\rho}$$

Identical weights can also be derived using a different definition for the ICC that assumes that there is an underlying latent variable that determines whether each individual experiences the outcome. This latent variable is assumed to follow the logistic distribution [16].

Weighting by cluster size and within-cluster variance have also been used elsewhere but are not considered here. Weighting by cluster size ignores the non-linear impact of cluster size on the information provided by a cluster and has been shown to give biased effect estimates [11, 17]. Weighting by within-cluster variance ignores the between cluster element of the cluster log-odds variance and has been shown to give inflated type-one error unless the outcome is very common [17].

Generalised linear mixed effect models

GLMMs with a binomial distribution and logit link directly model the between cluster variation so that:

$$E[\boldsymbol{y}|\boldsymbol{X},\boldsymbol{u}] = \frac{exp(\boldsymbol{X\beta} + \boldsymbol{u})}{1 + exp(\boldsymbol{X\beta} + \boldsymbol{u})}$$

where $\boldsymbol{y}$ is a vector of outcomes of length nm_{ij}, $\boldsymbol{X}$ is a nm_{ij} x 2 matrix of fixed effect covariates consisting of a vector of ones and the trial arm assignment, $\boldsymbol{\beta}$ is a vector of fixed-effect parameters consisting of the intercept (β_0) and log odds ratio comparing control and intervention conditions (β_1), and $\boldsymbol{u}$ is a vector of random effects for clusters with elements $u_{ij} \sim N(0, \sigma_b^2)$.

Maximum likelihood is used to estimate the parameters $\boldsymbol{\beta}$, and σ_b^2. For CRTs with a continuous outcome, a similar GLMM is used, but with a normal distribution with identity link. Maximum likelihood estimation leads to biased estimation of σ_b^2 with a small number of clusters, and use of restricted maximum likelihood estimation, which applies a transformation to the data to remove fixed effects before estimating random effect variances, reduces this bias [18].

For the binary outcome model above, the marginal likelihood does not have a closed form expression, so numerical integration methods are required. Adaptive quadrature is a commonly used technique, but there is no equivalent restriction technique to the one used to reduce bias with continuous outcomes. Pseudo-likelihood [19] and penalised quasi-likelihood [20] perform less well than adaptive quadrature with a large number of clusters [20], but methods of restriction are available. Elff et al. found that this made penalised quasi-likelihood a more suitable technique for data with fewer clusters with a common outcome and probit link [5].

In addition to selection of integration method, confidence intervals and p-values should be constructed with a t-distribution when the number of clusters is small to account for uncertainty in estimation of the standard error. There are three commonly used options for the degrees of freedom:

- Clusters minus cluster-level parameters. In an unadjusted analysis, this is $DF_{C-P} = n - 2$
- Satterthwaite [21]. For a test of the intervention effect parameter, degrees of freedom are

$$DF_S = \frac{2Var(\beta_1)^2}{Var[Var(\beta_1)]}$$

Where $Var[Var(\beta_1)]$ is approximated using the multivariate delta method.

- Kenward and Roger [22] developed a small sample correction that involves inflation of the standard error as well as a degree of freedom correction ($\mathrm{DF}_{\mathrm{KR}}$). For a test of the intervention effect parameter, the degrees of freedom part of this correction are the same as DF_{S}, but the standard error may be different.

With continuous outcomes, Satterthwaite degrees of freedom give closest to nominal type-one error, and clusters minus cluster-level parameters and Kenward-Rogers degrees of freedom are more conservative [11]. With binary outcomes, clusters minus cluster-level parameters has previously given closest to nominal coverage [6] with a common outcome.

With a large number of clusters, GLMM provide similar results with normally and non-normally distributed cluster effects (here these are cluster log-odds) except for extreme cases of non-normality or very large between cluster variability [12, 23]. Supporting Fig. 1B and C shows examples of cluster effect distributions that have not affect GLMM performance. However, this has not been studied for settings with a small number of clusters.

Generalised Estimating Equations

GEE model the marginal individual level data treating the correlation parameters as nuisance parameters. We assume a correlation structure for the data, which gives a covariance matrix $\boldsymbol{V}_{\boldsymbol{W}ij}$ for the vector of outcomes in each cluster $\boldsymbol{y}_{ij}$. We use the logit link so that

$$\boldsymbol{\mu} = E[\boldsymbol{y}|\boldsymbol{X}] = \frac{exp(\boldsymbol{X\beta})}{1 + exp(\boldsymbol{X\beta})}$$

The (uncorrected) sandwich covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$\boldsymbol{V_s} = \boldsymbol{V_M}\left[\sum\nolimits_{j=0}^{1}\sum\nolimits_{i=1}^{n/2}\boldsymbol{D_{ij}}\boldsymbol{V}_{\boldsymbol{W}ij}^{-1}\widehat{\boldsymbol{Cov}}\left(\boldsymbol{y_{ij}}\right)\boldsymbol{V}_{\boldsymbol{W}ij}^{-1}\boldsymbol{D_{ij}}\right]\boldsymbol{V_M}$$

where $\boldsymbol{V_M}$ is the model based variance of $\boldsymbol{\beta}$, $\boldsymbol{D}_{ij} = \partial\boldsymbol{\mu_{ij}}/\partial\boldsymbol{\beta'}$, and $\widehat{Cov}(\boldsymbol{y_{ij}}) = (\boldsymbol{y_{ij}} - \hat{\boldsymbol{\mu}}_{ij})(\boldsymbol{y_{ij}} - \hat{\boldsymbol{\mu}}_{ij})^T$.

With fewer than 40–50 clusters [24], the sandwich covariance estimator used in conjunction with GEE are known to estimate standard errors that are too small on average, hence they are negatively biased, and several bias corrections have been suggested to reduce this bias [7–9, 25, 26]. Corrections developed by Kauermann and Carroll [8] and Fay and Graubard [7] have had particularly promising performance across a range of scenarios where CRTs are used [10, 27]. Others were often conservative [9] or highly variable [26].

Kauermann and Carroll suggested the estimator: [8]

$$\boldsymbol{V_{KC}} = \boldsymbol{V_M}\left[\sum\nolimits_{j=0}^{1}\sum\nolimits_{i=1}^{n/2}\boldsymbol{D_{ij}}\boldsymbol{V}_{\boldsymbol{W}ij}^{-1}\boldsymbol{A_{KCij}}\widehat{\boldsymbol{Cov}}(\boldsymbol{y_{ij}})\boldsymbol{A_{KCij}}\boldsymbol{V}_{\boldsymbol{W}ij}^{-1}\boldsymbol{D_{ij}}\right]\boldsymbol{V_M}$$

where $\boldsymbol{A_{KCij}} = \left[\boldsymbol{I_{ij}} - \boldsymbol{D_{ij}}\boldsymbol{V_M}\boldsymbol{D_{ij}^T}\boldsymbol{V}_{\boldsymbol{W}ij}^{-1}\right]^{-1/2}$.

Fay and Graubard suggested the estimator: [7]

$$\boldsymbol{V_{FG}} = \boldsymbol{V_M}\left[\sum\nolimits_{j=0}^{1}\sum\nolimits_{i=1}^{n/2}\boldsymbol{A_{FGij}}\boldsymbol{D_{ij}}\boldsymbol{V}_{\boldsymbol{W}ij}^{-1}\widehat{\boldsymbol{Cov}}(\boldsymbol{y_{ij}})\boldsymbol{V}_{\boldsymbol{W}ij}^{-1}\boldsymbol{D_{ij}}\boldsymbol{A_{FGij}}\right]\boldsymbol{V_M}$$

where $\boldsymbol{A_{FGij}} = diag\left[\left(1 - min\left[0.75, \boldsymbol{D_{ij}}\boldsymbol{V}_{\boldsymbol{W}ij}^{-1}\boldsymbol{D_{ij}^T}\boldsymbol{V_M}\right]\right)\right]$

As well as standard error corrections, it is necessary to construct confidence intervals and p-values from a t-distribution as has been seen for cluster-level analysis and GLMM. There is less literature for GEE, but $\mathrm{DF}_{\mathrm{C\text{-}P}}$ has performed well [10].

Unlike the cluster level analysis and GLMM, GEE do not make any assumptions about the distribution of the cluster log-odds. They were developed as a robust method of analysis for non-normal outcomes, so we expect this method to be robust to non-normality of cluster log-odds.

Comparison between these methods

There is particularly sparse literature comparing the different types of methods. Since it may now be possible to maintain type-one error with a small number of clusters for all approaches, comparisons of power are relevant. While with a large number of clusters, it is known that the individual-level methods we've described are more powerful than the cluster-level method when clusters vary in size [28], there are few comparisons of power with small sample corrections applied to the individual-level methods. For continuous outcomes and normally distributed cluster-level summaries, GLMMs with a small sample correction had a higher power than unweighted cluster-level analyses and GEEs, and a similar power as inverse variance weighted cluster-level analyses [11]. Others have found that GLMM had higher power than GEE after small sample corrections had been applied [29].

Simulation study methods

We conducted a simulation study to compare the performance of the analysis methods described above. The simulations were performed in SAS software, version 9.4 of the SAS system for windows[30]. The scenarios included in the simulations are given in table 1 and more details are in the supporting information text. All combinations of each scenario were simulated with 1000 repetitions so that there is a 95% probability that true type-one error of 5% is estimated to be between 3.6% and 6.4%.

Data generating mechanism

We simulated binary, clustered data using the data generating model:

$$Y_{ij} \sim Binomial\left(\frac{exp\left[\beta_0 + \beta_1 i + u_{ij}\right]}{1 + exp\left[\beta_0 + \beta_1 i + u_{ij}\right]}, m_{ij}\right)$$

where Y_{ij} is the number of events in arm i in cluster j, β_0 is the true log-odds of the outcome in the control condition, β_1 is the log odds ratio intervention effect, and u_{ij} is a random effect for cluster with mean zero and variance σ_b^2.

The prevalence of the outcome in the control condition was either 10% or 30%. We simulated scenarios with and without an intervention effect (β_1). For the scenarios with an intervention effect, we used the Stata 15 [31] power command to select cluster-specific odds ratios that would be expected to have 80% power for each scenario. This command uses the design effect $1 + (m - 1)\rho$ to account for clustering and the design effect of van Breukelen, Candel, and Berger to account for unequal cluster size [32].

The ICC was set to 0.001, 0.01, 0.5, or 0.1 on the log odds scale, defined as $\sigma_b^2/(\sigma_b^2 + \pi^2/3)$, to span a range of common values in health research [33–35]. For the distribution u_{ij}, we considered a normal distribution, a uniform distribution to explore the impact of kurtosis, and a gamma distribution with shape parameter $\lambda = 2$. These distributions were selected as they are the limit for which GLMMs estimate unbiased cluster-level coefficients with a large number of clusters [12, 23].

Trial designs

We simulated trials with a total of 8, 12, 20, or 30 clusters and a 1:1 randomisation ratio. Cluster size was either common to all clusters or simulated to vary between clusters. Variable cluster sizes were drawn from a negative binomial distribution to give a minimum cluster size of 3 and coefficient of variation in cluster size of 0.5 (the median CV of UK primary care trust size [36]), or 0.8 (a large variability) [11, 37]. The mean cluster size was either 10, 50, or 1000 to represent small, medium and very large clusters [1].

Estimand and Analysis Methods

The estimands of interest for the analysis of the simulated trials was the odds ratio intervention effect and

Table 1 Summary of simulation study scenarios

Parameter	Number of scenarios	Values
Total number of clusters	4	8, 12, 20, 30
Mean Cluster size ($\overline{m}$)	3	10, 50, 1000
Coefficient of variation (CV) of cluster size	3	$CV = \frac{s}{\overline{m}} = 0, 0.5, 0.8$ Where s is the standard deviation in cluster sizes and$\overline{m}$is the mean cluster size, Cluster size $m_{ij}$is sampled from a negative binomial distribution as follows: $\delta \sim Negbin\left(nooffails = \frac{(\overline{m}-2)^2}{s^2-(\overline{m}-2)}, poffail = \frac{\overline{m}-2}{s^2}\right)$ $m_{ij} = 2 + \delta$
Control cluster prevalence	2	10%, 30%
Intervention effect	2	No effect, or odds ratio between 1.12 and 11.49 selected for each scenario to achieve 80% power
ICC	4	0.001, 0.01, 0.05, 0.1
Cluster effect distribution	3	Normal: $u_{ij} \sim N(0, \sigma_b^2)$ Gamma $u_{ij} = \frac{\sigma_b(a_{ij}-2)}{\sqrt{2}}$where $a_{ij} \sim Gamma(2, 1)$ Uniform $u_{ij} \sim Uniform\left(-\sqrt{3\sigma_b^2}, \sqrt{3\sigma_b^2}\right)$ Distributions are defined to give the specified between cluster variability set by the ICC

the statistical test for no intervention effect. We use a cluster specific estimand for the cluster-level analysis and GLMM and a population-averaged estimand for the GEE.

We analysed the data with all methods previously described: unweighted cluster analysis (CL-UNW), and inverse variance weights with equal variance within clusters (CL-W); GLMM using adaptive Gauss-Hermite quadrature (AQ) or restricted pseudo-likelihood (REPL) and degrees of freedom as clusters minus cluster-level parameters (DF_{CP}), Satterthwaite (DF_S), or Kenward-Rogers (DF_{KR}) where DF_S and DF_{KR} were only available for REPL; and GEE with a sandwich variance with the Kauermann and Carroll correction (KC) or Fay and Graubard correction (FG) with boundary parameter 0.75 using DF_{CP} and an independent (I) or exchangeable (E) working correlation matrix. The exchangeable working correlation matrix GEE were only run on scenarios with mean cluster size of 10 or 50 due to unfeasibly long run time with a cluster size of 1000.

Our data generating mechanism specified a cluster-specific intervention effect odds ratio. Since GEE estimate population-averaged (marginal) odds ratios, we estimated the true marginal effects to compare to GEE estimates using the approximate formula [38]:

$$\beta_{Marginal} \approx \beta_{Conditional}\left(\left[\frac{16\sqrt{3}}{15\pi}\right]^2 \sigma_b^2 + 1\right)^{-1/2}$$

First we select the cluster-level, GLMM, and GEE method that has most consistently controls type-one error in the simulation study; then we compare the performance of these three methods.

Performance measures

For each scenario and analysis method, we calculated the standardised bias of the intervention effect estimate, which is the bias as a percentage of the standard deviation of the intervention effect estimates across the 1000 repetition; relative bias in standard errors; type-one error; and power [39]. We also calculated coverage but results are not shown due to the similarity of results to type-one error results. We refer to type-one error less than 3.6% as conservative and type-one error above 6.4% as inflated. Convergence rates were summarised and analyses that did not converge were excluded.

Results

For each type of analysis, we will summarise important results, more detail is given in the supporting information.

Intervention effect estimate bias, standard error bias, and type-one error

Cluster-level methods

Less than 1% of CL-UNW or CL-W failed to provide intervention effects or p-values in all scenarios. Non-convergence only occurred if all clusters-level summaries were identical within each arm, giving a within-arm variance of zero.

CL-UNW and CL-W had mean intervention effect estimate standardised bias of 52% and 38% closer to the null respectively across scenarios with mean cluster size 10 and low outcome prevalence (Fig. 1). Both methods demonstrated no bias in any other scenario (mean standardised bias -2% for CL-UNW and -1% for CL-W).

Standard errors of CL-UNW were within 10% of the standard deviation of simulated estimates in all scenarios, and type-one error was close to nominal in all scenarios (Fig. 1). Confidence interval coverage was often low with mean cluster size 10 and low outcome prevalence (102/144(71%) scenarios had coverage < 93.6%) due to the biased intervention effect estimate.

CL-W standard errors were between 46% smaller and 6% larger than the standard deviation of estimates with mean cluster size 10 and low prevalence. Standard errors were closer to the standard deviation of simulated estimates in other scenarios (between 9% smaller and 14% larger). Type-one error for CL-W was inflated with mean cluster size 10 and low outcome prevalence (97/144(67%) scenarios with type-one error > 6.4%) and when cluster size varied (type-one error > 6.4% in 25/240(10%) scenarios with cluster size CV = 0.5, and 49/240(20%) with cluster size CV = 0.8).

Supporting Figs. 2, 3, 4, 5 and 6 show cluster-level analysis performance by each simulation study parameter.

GLMM

In both REPL and AQ, up to 10% of models failed to converge with 8 clusters with mean cluster size of 10 and low outcome prevalence. REPL resulted in up to 8% non-convergence with mean cluster size 1000 and ICC = 0.001; this was more pronounced with 30 clusters but persisted with 20 clusters. Non-convergence was less than 5% in all other scenarios.

Both REPL and AQ gave estimates of the intervention effect with minimal bias in all scenarios (mean 2.9% standardised bias for AQ and 0.6% for REPL across all scenarios, Fig. 2).

AQ resulted in standard errors that were too small with mean cluster size 1000 with 20 or fewer clusters (mean 8% smaller than the standard deviation of simulations with 12 clusters and 12% with 8 clusters). Bias in standard errors increased with larger ICC (ICC = 0.1 standard error bias = -6%, ICC = 0.001, standard error bias = -3%).

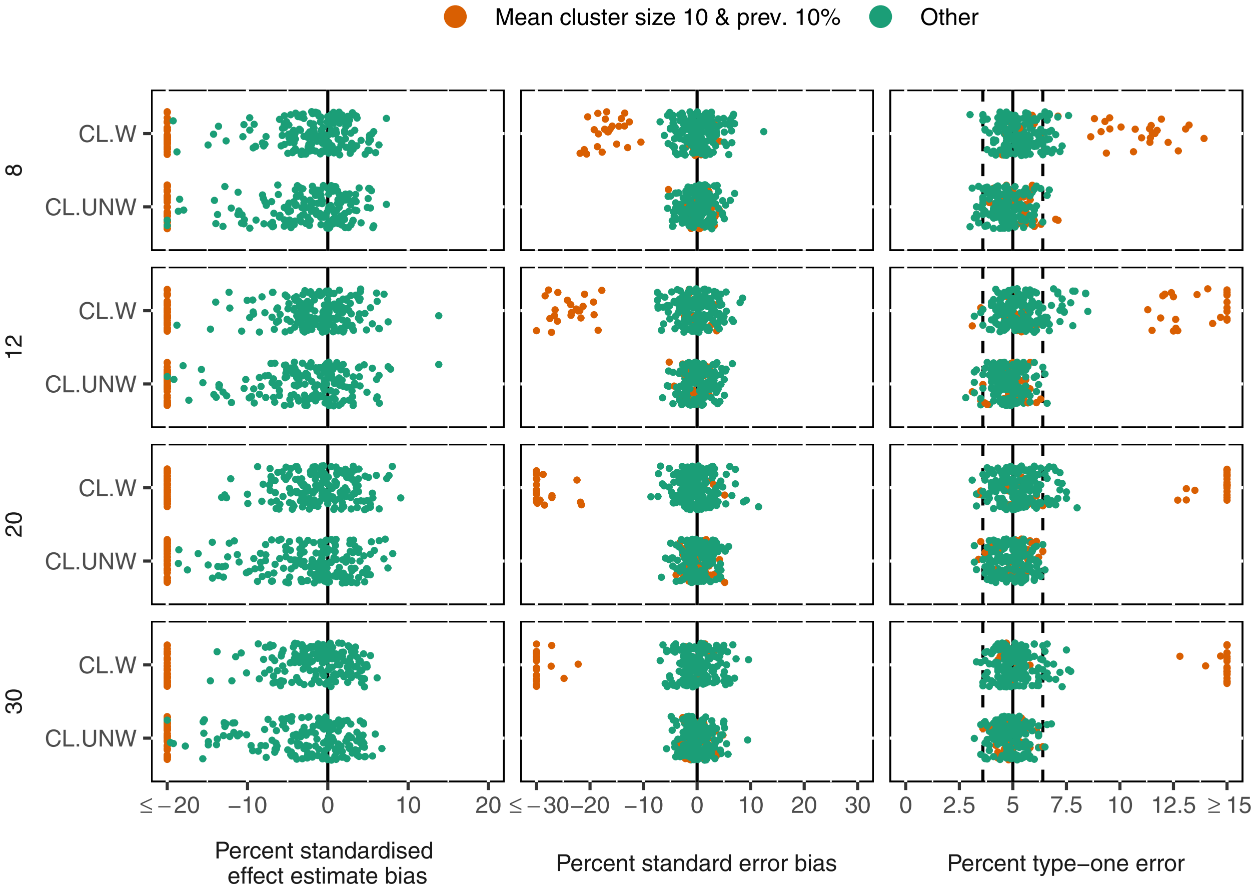

Fig. 1 Performance measures of cluster-level analysis methods by number of clusters (rows), cluster size and outcome prevalence (colour). Measures shown (columns): Standardised intervention effect estimate bias, standard error bias, type-one error. Each dot represents a scenario summarised over the 1000 repetitions. All 864 scenarios are shown for each measure

This led to mean type-one error of 7% with ICC = 0.1, mean cluster size of 1000 and 8 or 12 clusters.

REPL gave more consistent standard errors. However, with mean cluster size 10, standard errors were a mean 4% larger than the standard deviation of simulations with 12 cluster and 10% larger with 8 clusters. The Kenward-Roger standard error correction had minimal impact. Combining REPL with DF_{CP} controlled type-one error, but was conservative in the scenarios where the standard errors were inflated (mean cluster size 10 had mean type-one error = 3.8% with 30 clusters, 1.9% with 12 clusters, and 0.8% with 8 clusters). DF_{KR} and DF_{SA} were more conservative.

Supporting Figs. 7, 8, 9 and 10 show GLMM methods performance by each simulation study parameter.

GEE

With an independent working correlation matrix, in most scenarios, less than 3% failed to converge. Up to 9% failed to converge with mean cluster size 10 and outcome prevalence 10%. With an exchangeable working correlation matrix, non-convergence was common with varying cluster size. With cluster size CV = 0.8 and mean cluster size 10, a mean of 20% failed to converge; with mean cluster size 50, this increased to a mean of 34% failing to converge.

The intervention effect was estimated with little bias where no effect was present and negligible bias (compared to estimated marginal effects) when the intervention did have an effect in truth (mean 5% and 3% standardised bias for independent and exchangeable working correlation matrices respectively, Fig. 3). Variability of effect estimates was similar between the independent and exchangeable working correlation matrices.

With an independent working correlation matrix, both KC and FG standard errors demonstrated little bias with 20 or more clusters. With 12 or fewer clusters, FG standard errors were on average 6% too large and KC standard errors were on average 2% too small (Fig. 3). With FG standard errors, type-one error was conservative in 25% of scenarios and inflated in 3% of scenarios. Inflated

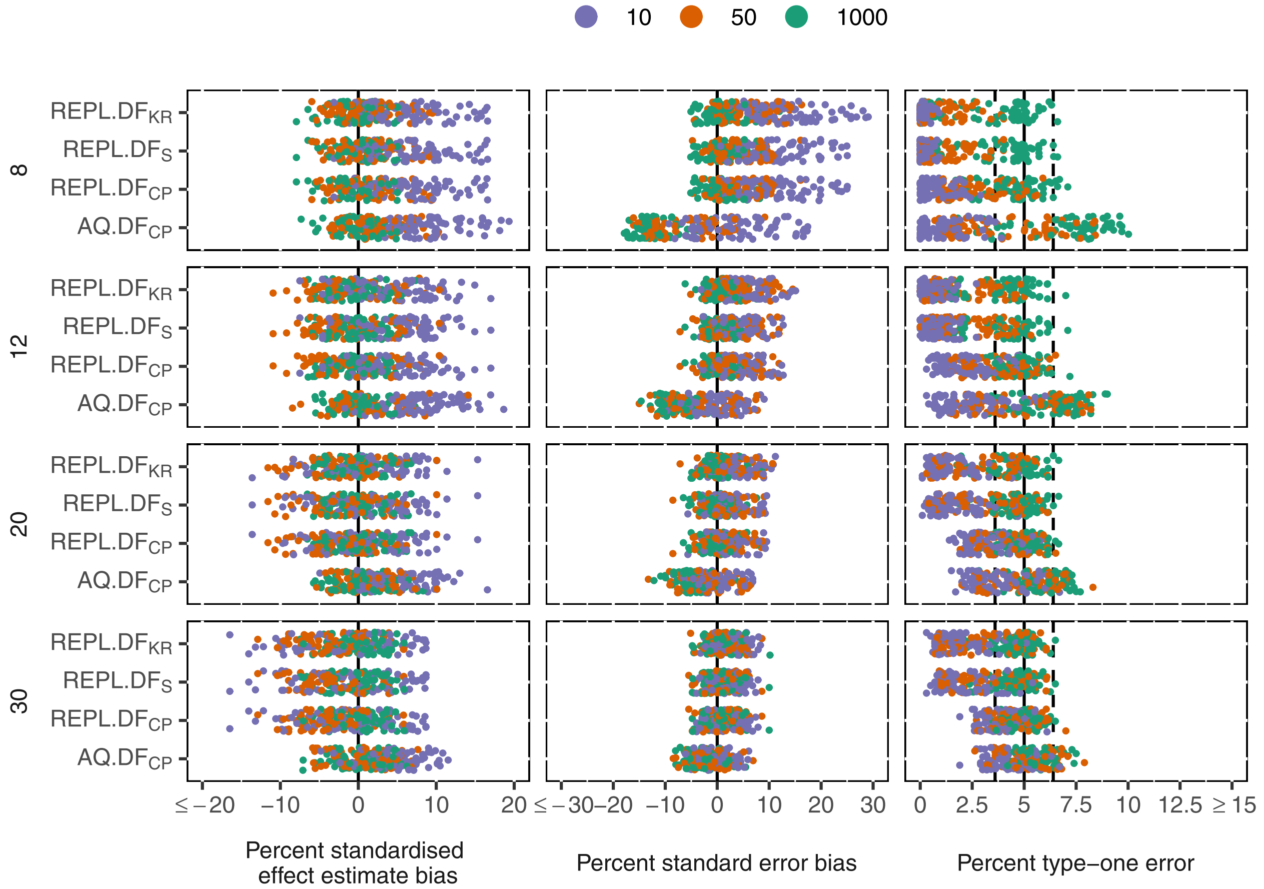

Fig. 2 Performance measures of GLMM methods by number of clusters (rows), and mean cluster size (colour). Measures shown (columns): Standardised intervention effect estimate bias, standard error bias, type-one error

type-one error occurred when clusters were large and there was large variability in cluster size.

With an exchangeable working correlation matrix, FG standard errors became more variable and had a mean 10% overestimation of standard errors with 12 or fewer clusters. KC standard errors had little bias with 12 or more clusters and were a mean 3% overestimated with 8 clusters. With FG standard errors, type-one error was conservative in 33% of scenarios and inflated in 1% of scenarios.

In order to select a working correlation matrix to take forward to compare with GLMM and CL methods, we also looked at power. With FG standard errors and DF$_{CP}$, power was similar with an independent or exchangeable working correlation matrix (mean 1% greater power with an independent working correlation matrix). Power was similar with 20–30 clusters (exchangeable mean 0.4% higher) but favoured an independent working correlation matrix with fewer cluster (exchangeable mean 0.8% and 3.8% lower with 12 and 8 cluster respectively). Power favoured an independent working correlation matrix with ICC = 0.001 (mean 3.1% higher power than exchangeable) but favoured an exchangeable working correlation matrix with ICC = 0.1 (mean 1.6% higher power with exchangeable).

Due to the similar power and better convergence, we carried an independent working correlation matrix forward to the comparison with the cluster-level method and GLMM.

Supporting Figs. 11, 12, 13 and 14 show GEE method performance by each simulation study parameter.

Comparison of cluster-level method, GLMM, and GEE

Next we compare the best performing method from each analysis type: CL-UNW, REPL.DF$_{CP}$, and GEE with FG standard errors and an independent working correlation matrix (FG.I.DF$_{CP}$).

All three method controlled type-one error in the majority of scenarios (Fig. 4). CL-UNW controlled type-one error most consistently; only 1% of scenarios had inflated and 5% conservative type-one error. REPL-BW

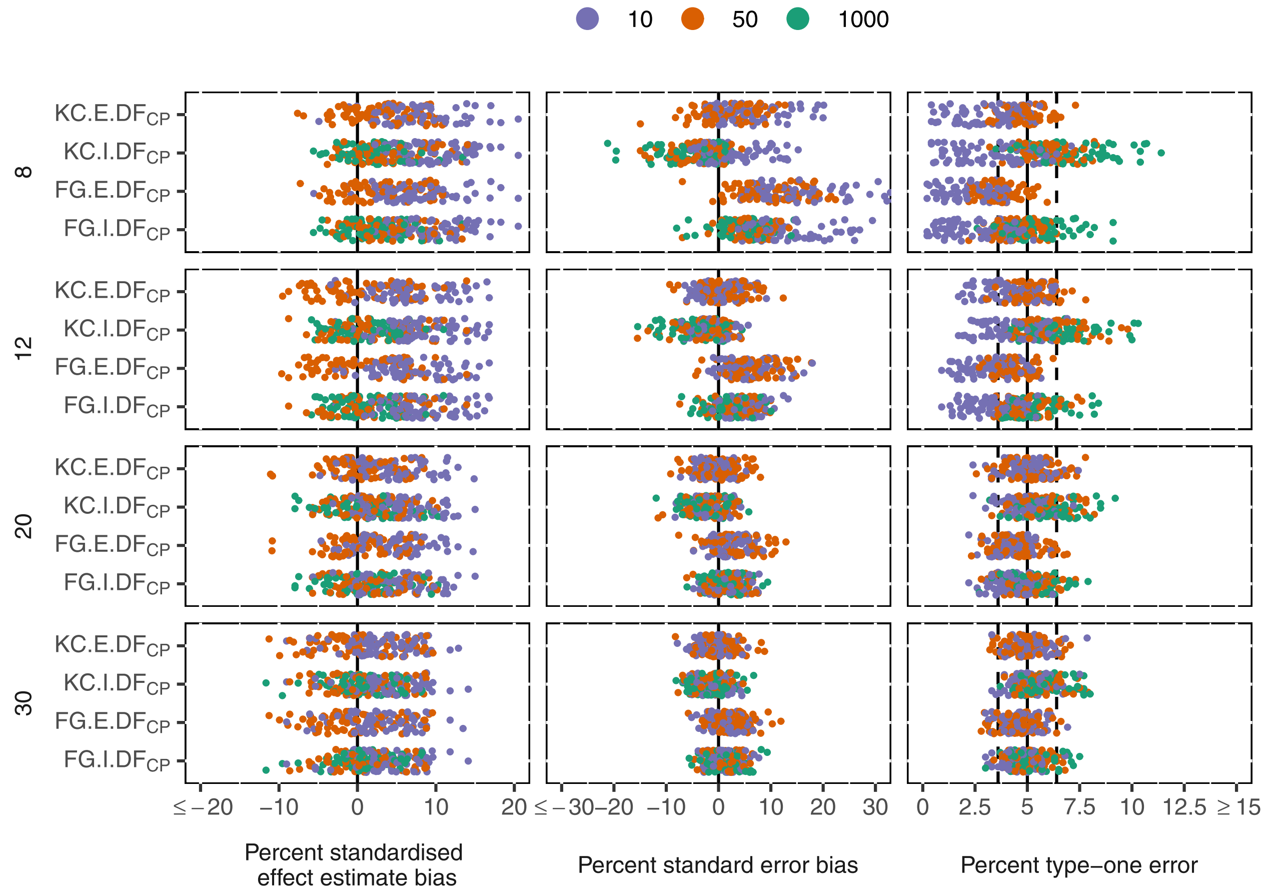

Fig. 3 Performance measures of GEE methods by number of clusters (rows), and mean cluster size (colour). Measures shown (columns): Standardised intervention effect estimate bias, standard error bias, type-one error

has the most conservative type-one error: 42% of scenarios had conservative type-one error and 1% has inflated type-one error. FG.I.DF$_{CP}$ had the most variable type-one error: 6% of scenarios had inflated type-one error and 17% conservative.

Power

Excluding scenarios with low prevalence and cluster size of 10 (due to biased effect estimates from CL-UNW), REPL.DF$_{CP}$ has on average 2% higher power than CL-UNW, and FG.I.DF$_{CP}$ had on average 3% lower power than CL-UNW (Fig. 5).

The power difference between CL-UNW and FG.I.DF$_{CP}$ was most strongly influenced by the ICC with higher ICC leading to CL-UNW having higher power: with ICC = 0.1 CL-UNW had a mean 10% higher power than FG.I.DF$_{CP}$, but with ICC = 0.001, FG.I.DF$_{CP}$ had a mean 8% greater power than CL-UNW. With fewer clusters, power favoured CL-UNW.

The power difference between REPL.DF$_{CP}$ and CL-UNW was most strongly influenced by variation in cluster size. With common cluster size, CL-UNW had a mean 2% higher power than REPL.DF$_{CP}$; with CV = 0.5, REPL-BW had a mean 2% higher power than CL-UNW; and with CV = 0.8, REPL.DF$_{CP}$ had a mean 10% higher power than CL-UNW. With fewer clusters, power became more similar between the methods. With higher ICC, power become more similar between the methods.

Across all scenarios (including scenarios with low prevalence and cluster size of 10), REPL.DF$_{CP}$ had a mean 5% higher power than FG.I.DF$_{CP}$. This was most strongly influenced by ICC: with ICC = 0.1, REPL.DF$_{CP}$ had a mean 10% higher power, but with ICC = 0.001, REPL.DF$_{CP}$ and FG.I.DF$_{CP}$ had similar power. Number of clusters had minimal impact on the difference in power.

Supporting Figs. 15, 16 and 17 show power comparisons by each simulation study parameter.

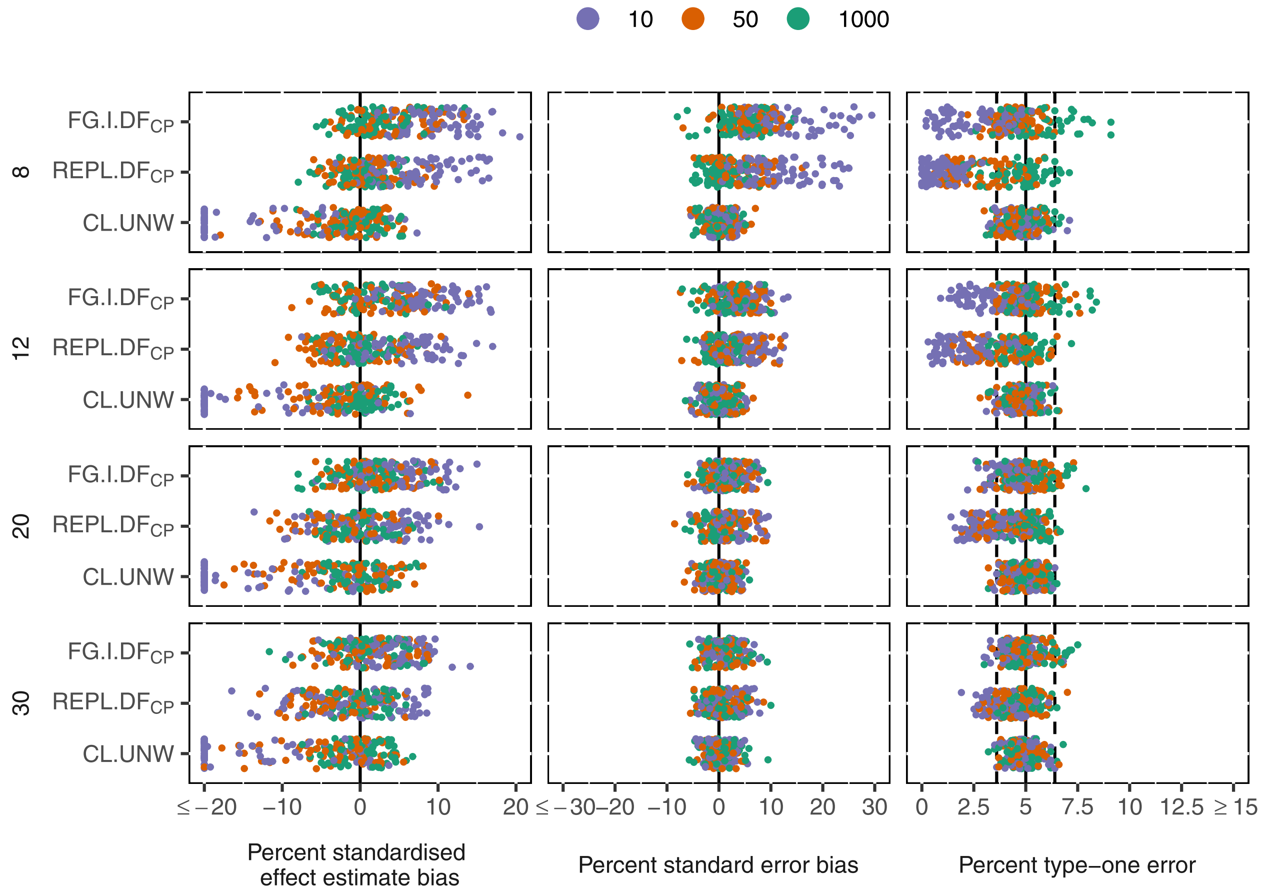

Fig. 4 Comparison of bias and type-one error of unweighted cluster-level analysis, GLMM with REPL and DF_{CP}, and GEE with FG standard errors and DF_{CP} by number of clusters (rows), and mean cluster size (colour). Measures shown (columns): Standardised intervention effect estimate bias, standard error bias, type-one error

Robustness to non-normality

There was no difference in our findings based on whether the clusters log-odds were distributed normally, or with skew or kurtosis (supporting Figs. 3, 8, 12, and 16).

Recommendations

Table 2 summarises our results to provide recommendations of the most robust and powerful analysis by scenario.

Illustrative example

Treatment for tuberculosis involves 6 months of alternate daily drugs. Recovery is hampered by non-adherence to treatment and the standard of care (SoC) is directly observed therapy where a health worker or family member directly observes the patient taking their medication. This is costly and has limited impact on adherence [40].

In one trial, two interventions, a text message reminder and electronic monitoring box for medication, were compared to SoC [41]. Here, we focus on the comparison of the monitoring box to SoC: a comparison with 9 clusters in each arm. Randomisation was stratified by province and whether clusters were urban: this is ignored for simplicity in this example. In the monitor box arm, patient's medication was stored in a box that recorded openings of the box that clinicians could review to assess the need for adherence counselling, and a light and sound reminded patients to take medication. We will focus on a secondary outcome from the trial: whether patients missed more than 10% of doses over treatment. There was a mean 116 patients per cluster with coefficient of variation 0.1. The ICC on the log-odds scale was estimated as 0.09 (estimated from the REPL GLMM).

Figure 6 shows the estimated intervention effects, confidence intervals, and p-values from each of the methods considered in this paper. The outcome was common in both arms: 59% and 41% in the control and intervention arms missed more than 10% of doses respectively.

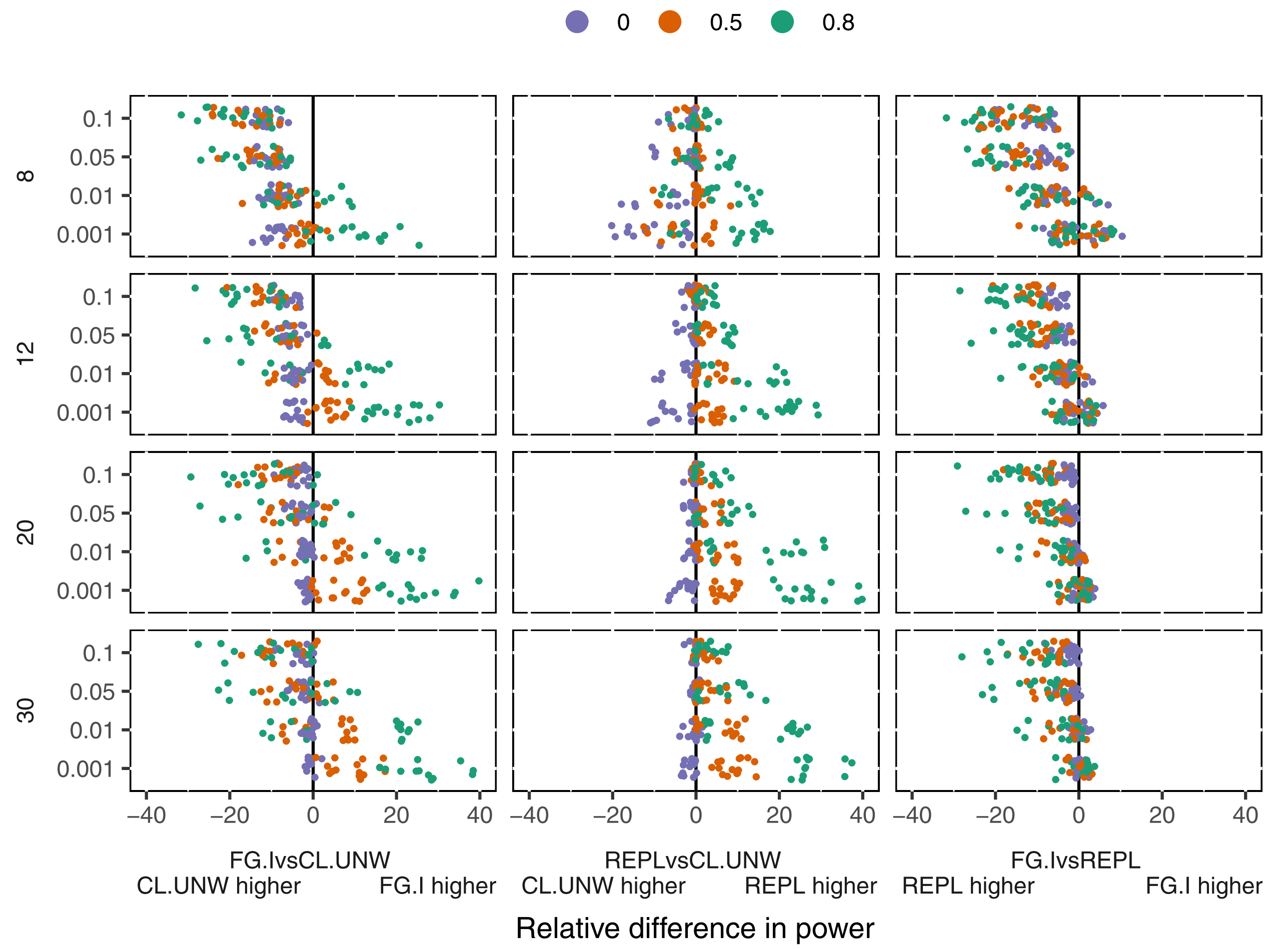

Fig. 5 Power comparison of unweighted cluster-level analysis (CL.UNW), GLMM with REPL and DF_{CP} (REPL), and GEE with FG standard errors and DF_{CP} (FG.I) (columns) by number of clusters (rows), ICC (y axis), and variability of cluster size (colour)

Table 2 Summary of simulation study results and recommendations on their use

	Cluster-Level Method	GLMM	GEE
Method	Use equal weighting of clusters and clusters minus cluster-level parameters degrees of freedom	Use restricted pseudo-likelihood and clusters minus cluster-level parameters degrees of freedom	Use Fay and Graubard standard errors, clusters minus cluster-level parameters degrees of freedom, and an independent working correlation matrix
Valid results [a]	Cluster size > 10 Or Common outcome (prevalence > 10%)	All scenarios	Cluster size ≤ 50 Or CV cluster size ≤ 0.5
Competitive power [b]	Common cluster size Or High ICC (ICC > 0.05)	Varying cluster size Or 20 + clusters	Low ICC ≤ 0.01

[a] Unbiased effects with controlled or conservative type-one error

[b] The method/s with greatest or similar to greatest power in a scenario

All analyses find strong evidence that the monitor box improved adherence compared to SoC, but the strength of evidence varied in line with the simulation study results.

Since clusters are medium sized, Table 2 recommends use of CL-UNW or $REPL.DF_{CP.}$ These provide almost identical results (CL-UNW: OR = 0.45 95% CI [0.25,

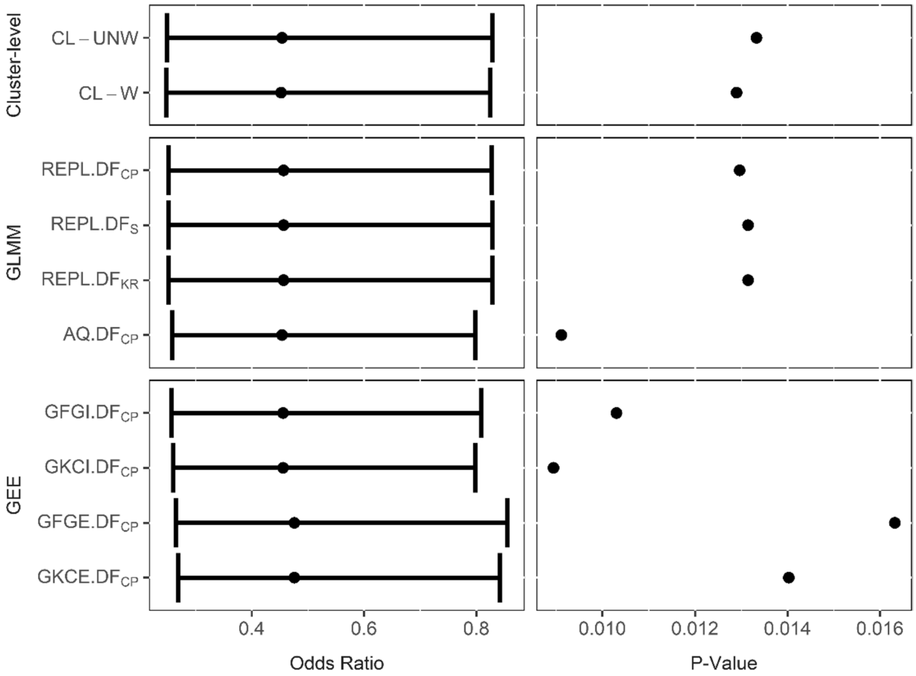

Fig. 6 Motivating example results analysed by all methods considered in the simulation study. Left panel shows odds ratios and confidence intervals, right panel shows *p* values. Rows are analysis methods

0.83] $p=0.013$, REPL.DF$_{CP}$: OR = 0.46 95% CI [0.25, 0.83] $p=0.013$).

Since this example has small variability in cluster size, CL-W would be likely to have nominal type-one error, hence the similar result to CL-UNW. The inflated type-one error of AQ.DF$_{CP}$ has resulted in a smaller *p*-value ($p=0.009$) and narrower confidence interval (0.26, 0.80), but the degrees of freedom has had little impact on REPL as the clusters are not small in this trial. GEE methods may have inflated type-one error due to the cluster size, but were also found to have lower power than other methods. This has led to more variable results ($p=0.009$ to 0.016).

Discussion

We have identified methods to control the type-one error with as few as 8 clusters with cluster-level analysis, GLMM, and GEE in high and low prevalence settings. Cluster-level analysis should give equal weight to all clusters. GLMM should use REPL to integrate the likelihood, and GEE should use the small sample standard error correction from Fay and Graubard. All methods require a t-distribution with clusters minus cluster-level parameters as the degrees of freedom to calculate confidence intervals and *p*-values. We found that unweighted cluster-level analysis had greatest power with common cluster size and competitive power when the ICC was high. GLMMs using REPL had greatest power with varying cluster size or 20 or more clusters despite conservative type-one error. GEE with FG standard errors tended to have equal or lower power than the other methods. All methods performed well with non-normally distributed cluster effects.

Our comparison of cluster-level analysis methods identified problems with inverse-variance weighting of clusters. The weighted least squares method assumes that the weights are known, when in truth they are estimates. This leads to the bias in standard errors we observed with CL-W [42]. Use of a robust standard error may be able to account for weight estimation, but this is likely to lead to similar results to the GEE methods shown in this paper, which had lower power. Cluster-level analysis is simple to code manually in any software, and a user written Stata command clan is also available [43].

We found that the integration method used in the GLMM was important for obtaining unbiased standard errors: REPL outperformed AQ. This extends the findings of Elff et al [5] to logistic models and to low prevalence outcomes. We used the SAS glimmix procedure to implement this method. The R function glmmPQL implements a similar method [44]. To our knowledge, REPL is not available in the statistical software Stata. Our findings of nominal type-one error with clusters minus cluster-level parameters and conservative type-one error with Kenward-Rogers and Satterthwaite are supported by previous research [6], and we determine that these findings hold for a low prevalence outcome.

For GEE, our recommendation of FG standard errors with clusters minus cluster-level parameter degrees of freedom are in agreement with others [10, 45–47]. This previous literature has not reported rates of convergence, which we found were low in some scenarios with an exchangeable working correlation matrix. The poor convergence was likely due to computational complexity from inversion of a matrix the same size as the cluster size and some iterations requiring inversion of a not invertible matrix [48]. Our comparison of power by choice of working correlation matrix is novel. With a large number of clusters, correct choice of working correlation matrix is known to improve power [49]. We found that this difference diminished with a small number of clusters so that there was little benefit from fitting the more complex exchangeable working correlation matrix. Similar results have been seen for stepped wedge cluster randomised trials [50] with a cluster size of 50 or less, so our finding of similarity is likely do to the smaller cluster sizes used for this comparison. This method is widely implemented in statistical software: we used the glimmix procedure in SAS [30], in Stata a user written command xtgeebcv is available [51], and in R the saws package implements the FG correction [7].

Our comparison of power from the three types of analysis is novel. We found that despite conservative type-one error, GLMM generally had greater or competitive power compared to GEE or cluster-level analysis. Contrary to settings with a large number of clusters, we found that cluster-level analysis maintained competitive power when the ICC was large, even with varying cluster size. GEE had competitive power with a low ICC, but often lower power than GLMM. Low power was also identified by Leyrat et al. for continuous outcomes [11]. Where power and type-one error are similar between the methods, and convergence of GEE is reasonable, the choice could be guided by whether researchers are interested in estimating a cluster-specific or population-averaged intervention effect. Another consideration of this choice is non-normality of the cluster effects. We found no difference in the performance of methods with the distributions we considered. These were distributions where the limits of where mixed effect models perform well with a large number of clusters [12], so if a larger degree of non-normality is suspected, we recommend use of GEE or cluster-level analysis.

Our findings were similar regardless of the distribution of cluster log-odds for all numbers of clusters that we considered. Since our choice of non-normality was the boundary of good GLMM performance with a large number of clusters, this suggests a similar performance of methods with a small number of clusters to their performance with a large number of clusters [12, 23]. Therefore, a cluster-level analysis should be used with very skewed data, but either method remains suitable with some skew or data that shows kurtosis.

All our selected methods struggled with a mean 10 observations per cluster and low outcome prevalence, but this scenario of few clusters, that are small, and with low prevalence is unlikely to occur in practice. Unweighted cluster-level analysis gave biased intervention effect estimates in these scenarios due to the presence of clusters with no events. GLMM and GEE methods overestimated standard errors in these scenarios. With larger clusters, low prevalence generally had little impact on results.

Our simulation study covered a broad range of scenarios where CRTs are common. However, there are analysis methods that we did not consider, which could have superior performance. This includes other small sample corrections available for GEE [10, 25, 27, 52]; some of these could have improve the type-one error control of GEE particularly for scenarios with large variability in cluster size [25], an over-dispersed binomial model [53, 54], and non-parametric methods such as permutation tests. Our analysis methods have not adjusted for covariates, and our simulations used simple randomisation. The methods we have considered may all be impacted by adjustment for cluster-level covariates. We have only considered estimation of the intervention effect, which is a cluster-level covariate. The performance of GEE methods may have been impacted by comparison to estimated marginal effects because our data generating mechanism used cluster-specific effects. We excluded any runs that did not converge. This was very common for GEE, and may have biased the estimated power if runs were more or less likely to converge if they identified an intervention effect: to our knowledge, this was not the case. Further, we only considered prevalence as low an 10%.

Conclusion

We recommend that CRTS with 30 or fewer clusters and a binary outcome use an unweighted cluster-level analysis, or GLMM using REPL. Confidence intervals and p-values for both methods should be calculated based on a t-distribution with the number of degrees of freedom defined as the number of clusters minus cluster-level parameters.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-022-01699-2.

Additional file 1. Supporting Information.

Additional file 2. Code to replicate illustrative example.

Acknowledgements

Not applicable

Authors' contributions

RJH conceived the idea for this publication. JAT designed and conducted the simulation study, interpreted results and wrote the first draft of the manuscript. All authors provided input into the design and interpretation of the simulation study and read and approved the final manuscript.

Funding

JAT, KF, and RH are jointly funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement and is also part of the EDCTP2 programme supported by the European Union [grant nos MR/R010161/1 and MR/K012126/1]. CL is supported by the UK Medical Research Council (Skills Development Fellowship MR/T032448/1).

Availability of data and materials

Data for the illustrative example are available at http://doi.org/10.17037/DATA.4. Code to implement the methods discussed in this article is provided in the supplementary material.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors of this article have no competing interests to declare.

Author details

[1]Department of Infectious Disease, London School of Hygiene & Tropical Medicine, London, UK. [2]Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, UK.

Received: 31 August 2021 Accepted: 11 July 2022
Published online: 12 August 2022

References

1. Kahan BC, Forbes G, Ali Y, et al. Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. Trials. 2016;17:438. https://doi.org/10.1186/s13063-016-1571-2.
2. Hayes RJ and Moulton LH. Cluster Randomised Trials. New York: CRC Press; 2017.
3. Donner A, Klar N. Methods for comparing event rates in intervention studies when the unit of allocation is a cluster. Am J Epidemiol. 1994;140:279–89. https://doi.org/10.1093/oxfordjournals.aje.a117247 discussion 300-271. 1994/08/01.
4. Boneau CA. The effects of violations of assumptions underlying the t test. Psychol Bull. 1960;57:49–64. https://doi.org/10.1037/h0041412.
5. Elff M, Heisig P, Schaeffer M, et al. Multilevel Analysis with Few Clusters: Improving Likelihood-based Methods to Provide Unbiased Estimates and Accurate Inference. Br J Polit Sci. 2019;51(1):412–26. https://doi.org/10.1017/S0007123419000097.
6. Li P, Redden DT. Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. BMC Med Res Methodol. 2015;15:38. https://doi.org/10.1186/s12874-015-0026-x.journal article.
7. Fay MP, Graubard BI. Small-sample adjustments for Wald-type tests using sandwich estimators. Biometrics. 2001;57:1198–206. https://doi.org/10.1111/j.0006-341X.2001.01198.x.
8. Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. J Am Stat Assoc. 2001;96:1387–96. https://doi.org/10.1198/016214501753382309.
9. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics. 2001;57:126–34. https://doi.org/10.1111/j.0006-341X.2001.00126.x.
10. Li P, Redden DT. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. Stat Med. 2015;34:281–96. https://doi.org/10.1002/sim.6344 2014/10/28.
11. Leyrat C, Morgan KE, Leurent B, et al. Cluster randomized trials with a small number of clusters: which analyses should be used? Int J Epidemiol. 2018;47:321–31. https://doi.org/10.1093/ije/dyx169.
12. Litière S, Alonso A, Molenberghs G. The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. Stat Med. 2008;27:3125–44.
13. Ukoumunne OC, Carlin JB, Gulliford MC. A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials. Stat Med. 2007;26:3415–28. https://doi.org/10.1002/sim.2769.
14. Heeren T, D'Agostino R. Robustness of the two independent samples t-test when applied to ordinal scaled data. Stat Med. 1987;6:79–90. https://doi.org/10.1002/sim.4780060110 1987/01/01.
15. Kerry SM, Martin BJ. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. Stat Med. 2001;20:377–90.
16. Eldridge SM, Ukoumunne OC, Carlin JB. The Intra-Cluster Correlation Coefficient in Cluster Randomized Trials: A Review of Definitions. Int Stat Rev. 2009;77:378–94.
17. Westgate PM. On small-sample inference in group randomized trials with binary outcomes and cluster-level covariates. Biom J. 2013;55:789–806.
18. Patterson HD, Thompson R. Recovery of Inter-Block Information when Block Sizes are Unequal. Biometrika. 1971;58:545–54. https://doi.org/10.2307/2334389.
19. Wolfinger R, O'Connell M. Generalized linear mixed models a pseudo-likelihood approach. J Stat Comput Simul. 1993;48:233–43. https://doi.org/10.1080/00949659308811554.
20. Breslow NE, Clayton DG. Approximate Inference in Generalized Linear Mixed Models. J Am Stat Assoc. 1993;88:9–25. https://doi.org/10.2307/2290687.
21. Satterthwaite FE. An Approximate Distribution of Estimates of Variance Components. Biometrics Bulletin. 1946;2:110–4. https://doi.org/10.2307/3002019.
22. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics. 1997;53:983–97 1997/10/23.
23. Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. Biometrika. 2001;88:973–85. https://doi.org/10.1093/biomet/88.4.973.
24. Emrich LJ, Piedmonte MR. On some small sample properties of generalized estimating equation Estimates for multivariate dichotomous outcomes. J Stat Comput Simul. 1992;41:19–29. https://doi.org/10.1080/00949659208811388.
25. Ford WP, Westgate PM. Improved standard error estimator for maintaining the validity of inference in cluster randomized trials with a small number of clusters. Biom J. 2017;59:478–95. https://doi.org/10.1002/bimj.201600182 2017/01/28.
26. Morel JG, Bokossa MC, Neerchal NK. Small sample correction for the variance of GEE estimators. Biom J. 2003;45:395–409. https://doi.org/10.1002/bimj.200390021.
27. Lu B, Preisser JS, Qaqish BF, et al. A comparison of two bias-corrected covariance estimators for generalized estimating equations. Biometrics. 2007;63:935–41. https://doi.org/10.1111/j.1541-0420.2007.00764.x 2007/09/11.
28. Omar RZ, Thompson SG. Analysis of a cluster randomized trial with binary outcome data using a multi-level model. Stat Med. 2000;19:2675–88.

29. McNeish DM, Harring JR. Clustered data with small sample sizes: Comparing the performance of model-based and design-based approaches. Commun Stat Simul Comput. 2017;46:855–69. https://doi.org/10.1080/03610918.2014.983648.
30. SAS software version 9.4. SAS Institute Inc. Cary, NC. https://www.sas.com/images/contact_box/ReferenceFormatsforSASMaterials.pdf.
31. StataCorp. Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC; 2017.
32. van Breukelen GJP, Candel MJJM, Berger MPF. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. Stat Med. 2007;26:2589–603. https://doi.org/10.1002/sim.2740.
33. Adams G, Gulliford MC, Ukoumunne OC, et al. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. J Clin Epidemiol. 2004;57:785–94. https://doi.org/10.1016/j.jclinepi.2003.12.013.
34. Gulliford MC, Adams G, Ukoumunne OC, et al. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. J Clin Epidemiol. 2005;58:246–51. https://doi.org/10.1016/j.jclinepi.2004.08.012.
35. Pagel C, Prost A, Lewycka S, et al. Intracluster correlation coefficients and coefficients of variation for perinatal outcomes from five cluster-randomised controlled trials in low and middle-income countries: results and methodological implications. Trials. 2011;12:151. https://doi.org/10.1186/1745-6215-12-151.
36. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. Int J Epidemiol. 2006;35:1292–300. https://doi.org/10.1093/ije/dyl129.
37. Zou G, Donner A. Confidence Interval Estimation of the Intraclass Correlation Coefficient for Binary Outcome Data. Biometrics. 2004;60:807–11. https://doi.org/10.1111/j.0006-341X.2004.00232.x.
38 Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. Biometrics. 1988;44:1049–60.
39. White IR. simsum: Analyses of simulation studies including Monte Carlo error. Stata J. 2010;10:369.
40. Khan M, Walley J, Witter S, et al. Costs and cost-effectiveness of different DOT strategies for the treatment of tuberculosis in Pakistan. Health Policy Plan. 2002;17:178–86. https://doi.org/10.1093/heapol/17.2.178.
41. Liu X, Lewis JJ, Zhang H, et al. Effectiveness of Electronic Reminders to Improve Medication Adherence in Tuberculosis Patients: A Cluster-Randomised Trial. PLoS Med. 2015;12:e1001876. https://doi.org/10.1371/journal.pmed.1001876.
42. Carroll RJ, Wu CFJ, Ruppert D. The Effect of Estimating Weights in Weighted Least Squares. J Am Stat Assoc. 1988;83:1045–54. https://doi.org/10.1080/01621459.1988.10478699.
43. CLAN: Stata module to perform cluster-level analysis of cluster randomised trials. Nash S, Thompson JA and Leurent B. Boston College Department of Economics 2020.
44. Ronnegard L, Shen X, Alam M. hglm: A Package for Fitting Hierarchical Generalized Linear Models. R J. 2010;2:20–8.
45. Wang M, Kong L, Li Z, et al. Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples. Stat Med. 2016;35:1706–21. https://doi.org/10.1002/sim.6817 2015/11/21.
46. Scott JM, deCamp A, Juraska M, et al. Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. Stat Methods Med Res. 2014;2017(26):583–97. https://doi.org/10.1177/0962280214552092.
47. McNeish D, Stapleton LM. Modeling Clustered Data with Very Few Clusters. Multivariate Behav Res. 2016;51:495–518. https://doi.org/10.1080/00273171.2016.1167008.
48. Hanley JA, Negassa A, Edwardes MDd. GEE Analysis of negatively correlated binary responses: a caution. Stat Med. 2000;19:715–22. https://doi.org/10.1002/(SICI)1097-0258(20000315)19:5%3c715::AID-SIM342%3e3.0.CO;2-T.
49. Liang KY, Zeger SL. Longitudinal Data-Analysis Using Generalized Linear-Models. Biometrika. 1986;73:13–22. https://doi.org/10.1093/biomet/73.1.13.
50. Thompson J, Hemming K, Forbes A, et al. Comparison of small-sample standard-error corrections for generalised estimating equations in stepped wedge cluster randomised trials with a binary outcome: a simulation study. Stat Methods Med Res. 2020;0:0962280220958735. https://doi.org/10.1177/0962280220958735.
51. Gallis JA, Li F, Turner EL. xtgeebcv: A command for bias-corrected sandwich variance estimation for GEE analyses of cluster randomized trials. Stand Genomic Sci. 2020;20:363–81. https://doi.org/10.1177/1536867x20931001.
52. Westgate PM. A bias correction for covariance estimators to improve inference with generalized estimating equations that use an unstructured correlation matrix. Stat Med. 2013;32:2850–8. https://doi.org/10.1002/sim.5709.
53. Williams DA. Extra-Binomial Variation in Logistic Linear Models. J Roy Stat Soc: Ser C (Appl Stat). 1982;31:144–8. https://doi.org/10.2307/2347977.
54. Westgate PM, Cheng DM, Feaster DJ, et al. Marginal modeling in community randomized trials with rare events: Utilization of the negative binomial regression model. Clin Trials. 2022;19:162–71. https://doi.org/10.1177/17407745211063479.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.